CLAIMS

What is claimed is:

1. A method for training a Chinese language model from Chinese inputs, comprising:

extracting unknown character strings from a set of Chinese inputs;

determining valid words from the unknown character strings by comparing frequencies of occurrence of the unknown character strings with frequencies of occurrence of individual characters of the unknown character string; and

generating a transition matrix of conditional probabilities for predicting a word given a context.

2. The method of claim 1, wherein the extracting the unknown character strings utilizes a preexisting Chinese dictionary.

3. The method of claim 1, wherein the transition matrix of conditional probabilities is generated based on n-gram counts generated from the Chinese inputs where $n \geq 1$.

4. The method of claim 3, wherein the n-gram counts include the counts of n-tuples of adjacent and non-adjacent words in the set of Chinese inputs.

5. The method of claim 3, wherein the n-gram counts include the number of occurrences of each n-word sequence.

6. The method of claim 1, wherein an unknown character string is determined to be a valid new character string if the frequency of occurrence of the unknown character string as compared with frequencies of occurrence of the individual characters of the unknown character string is greater than a predetermined threshold.

7. The method of claim 1, wherein the set of Chinese inputs includes at least one of user Chinese inputs and a set of Chinese documents.

8. The method of claim 7, wherein the set of Chinese inputs includes a set of user Chinese queries to a web search engine.

9. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium on which are stored instructions executable on a computer processor, the instructions including:

extracting unknown character strings from a set of Chinese inputs;

determining valid words from the unknown character strings by comparing frequencies of occurrence of the unknown character strings with frequencies of occurrence of individual characters of the unknown character string; and

generating a transition matrix of conditional probabilities for predicting a word string given a context.

10. A system for training a Chinese language model, comprising:

a segmenter configured to segment unknown character strings from a set of Chinese inputs;

a new word analyzer configured to determine valid words from the unknown character strings by comparing frequencies of occurrence of the unknown character strings with frequencies of occurrence of individual characters of the unknown character string; and

a Chinese language model training module configured to generate a transition matrix of conditional probabilities for predicting a word string given a context.

11. The system of claim 10, wherein the segmenter segments the unknown character strings utilizing a preexisting Chinese dictionary.

12. The system of claim 10, wherein the new word analyzer is further configured to generate n-gram counts from the Chinese inputs where $n \geq 1$ and to generate the transition matrix of conditional probabilities based on the n-gram counts.

13. The system of claim 12, wherein the n-gram counts include the counts of n-tuples of adjacent and non-adjacent words in the set of Chinese inputs.

14. The system of claim 12, wherein the n-gram counts include the number of occurrences of each n-word sequence.

15. The system of claim 10, wherein the new word analyzer is further configured to determine that an unknown character string is a valid new character string if the frequency of occurrence of the unknown character string as compared with frequencies of occurrence of the individual characters of the unknown character string is greater than a predetermined threshold.

16. The system of claim 10, wherein the set of Chinese inputs includes at least one of user Chinese inputs and a set of Chinese documents.

17. The system of claim 16, wherein the set of Chinese inputs includes a set of user Chinese queries to a web search engine.

18. A method for translating a pinyin input to at least one Chinese character string, comprising:

generating a set of character strings from the pinyin input, each character string having a weight associated therewith indicating a likelihood that the character string corresponds to the pinyin input, the generating includes utilizing a Chinese dictionary including words extracted from a set of Chinese inputs and a language model trained based on the set of Chinese inputs.

19. The method of claim 18, wherein the set of Chinese inputs includes at least one of user Chinese inputs and a set of Chinese documents.

20. The method of claim 19, wherein the set of Chinese inputs includes a set of user Chinese queries to a web search engine.

21. The method of claim 18, further comprising:

prior to the generating, filtering out non-alphabetic characters from the pinyin input and storing their respective positions within the pinyin input; and

after the generating, merging each of the character strings with the non-alphabetic characters in positions corresponding to their stored positions.

22. The method of claim 18, further comprising:

prior to the generating, identifying an ambiguous word in the pinyin input, the ambiguous word being selected from a database of n-grams that are valid both in non-pinyin and in pinyin; and

analyzing context words of the user input to selectively classify the pinyin input as non-pinyin and pinyin, wherein the generating is performed only if the pinyin input is classified as pinyin.

23. The method of claim 18, further comprising generating a plurality of pinyin candidates from the pinyin input, wherein the generating includes generating a set of character strings for each pinyin candidate.

24. The method of claim 18, further comprising sorting and ranking the set of character strings according to the likelihood that the pinyin input corresponds to the character string.

25. The method of claim 18, wherein the generating includes performing a Viterbi algorithm utilizing the Chinese dictionary including words extracted from the set of Chinese inputs and the language model based on the set of Chinese inputs.

26. The method of claim 18, further comprising:

performing a search for a character string selected by a user from the set of character strings.

27. The method of claim 18, wherein the search is a web search performed by a search engine.

28. The method of claim 18, further comprising:

extracting unknown character strings from the set of Chinese inputs;

determining valid words from the unknown character strings by comparing frequencies of occurrence of the unknown character strings with frequencies of occurrence of individual characters of the unknown character string to generate the Chinese dictionary, the dictionary includes a mapping of the words to their corresponding pinyin; and

generating the language model for predicting a word string given a context.

29. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium on which are stored instructions executable on a computer processor, the instructions including:

generating a set of character strings from the pinyin input, each character string having a weight associated therewith indicating the likelihood that the character string corresponds to the pinyin input, the generating includes utilizing a Chinese dictionary including words extracted from a set of Chinese inputs and a language model trained based on the set of Chinese inputs.

30. A system for translating a pinyin input to at least one Chinese character string, comprising:

a pinyin-word decoder configured to generate a set of character strings from the pinyin input, each character string having a weight associated therewith indicating the likelihood that the character string corresponds to the pinyin input, the pinyin-word decoder being further configured to utilize a Chinese dictionary that includes words extracted from a set of Chinese inputs and a language model trained based on the set of Chinese inputs.

31. The system of claim 30, wherein the set of Chinese inputs includes at least one of user Chinese inputs and a set of Chinese documents.

32. The system of claim 30, further comprising a pinyin candidate generator configured to generate a plurality of pinyin candidates from the pinyin input, wherein the pinyin-word decoder is configured to generate a set of character strings for each pinyin candidate.

33. The system of claim 30, further comprising a sorting and ranking module configured to sort and rank the set of word strings according to the likelihood that the pinyin input corresponds to the character string.

34. The system of claim 30, wherein the pinyin-word decoder is further configured to execute a Viterbi algorithm utilizing the Chinese dictionary including words extracted from the set of Chinese inputs and the language model based on the set of Chinese inputs.

35. The system of claim 30, further comprising:

a segmenter configured to segment unknown character strings from the set of Chinese inputs;

a new word analyzer configured to determine valid words from the unknown character strings by comparing frequencies of occurrence of the unknown character strings with frequencies of occurrence of individual characters of the unknown character string; and

a Chinese language model training module configured to generate a transition matrix of conditional probabilities for predicting a word string given a context.

36. An pinyin classifier for classifying a user input, comprising:

a database of words that are valid both in non-pinyin and in pinyin; and

a classification engine configured to identify an ambiguous word in the user input selected from the database of words and to analyze context words of the user input to selectively classify the user input as non-pinyin or as pinyin.

37. The pinyin classifier of claim 36, wherein the classification engine is further configured to compute likelihoods of possible Chinese queries that may be generated from ambiguous query and to classify the user input as pinyin input if at least one of the likelihoods computed is above a predetermined threshold.

38. The pinyin classifier of claim 37, wherein the classification engine is configured compute the likelihoods of possible Chinese queries if the user input is unresolved after the classification engine analyzes the context words.

39. The pinyin classifier of claim 36, wherein the database of words that are valid both in non-pinyin and in pinyin is extracted from commonly occurring words in non-pinyin user queries.

40. A method for pinyin classification of a user input, comprising:

identifying an ambiguous word in the user input, the ambiguous word being selected from a database of n-grams that are valid both in non-pinyin and in pinyin; and

analyzing context words of the user input to selectively classify the user input as non-pinyin or as pinyin.

41. The pinyin classification method of claim 40, further comprising:

computing likelihoods of possible Chinese queries that may be generated from ambiguous query; and

classifying the user input as pinyin input if at least one of the likelihoods computed is above a predetermined threshold.

42. The pinyin classification method of claim 41, wherein the computing and classifying are performed if the user input is unresolved after the analyzing.

43. The pinyin classification method of claim 40, wherein the database of words that are valid both in non-pinyin and in pinyin is extracted from commonly occurring words in non-pinyin user queries.

44. A method for presenting possible translations of a user input, comprising:

providing a hyperlink for each possible translation of the user input, the user input and each possible translation of the user input being in different languages or language formats.

45. The method for presenting possible translations of claim 44, wherein the user input is in pinyin and each of the possible translations is in Hanzi.

46. The method for presenting possible translations of claim 44, further comprising:

providing at least one other hyperlink corresponding to a spelling correction of the user input.

47. The method for presenting possible translations of claim 44, wherein the hyperlink is to a web search of the corresponding possible translation of the user input.